International Comparisons of Student Achievement in Mathematics and Science: A Canadian Perspective

Philip Nagy
ontario institute for studies in education of the university of toronto

International testing programs in mathematics and science require careful interpretation. The major difficulties in such tests include test content substructure, the meaning of composite scores, the difficulty of measuring opportunity-to-learn, and the choice of test content. Valid interpretations must contend with inherent limitations to international comparisons, such as differing value systems, the differences among countries, and differences in enrolment rates. Opportunity-to-learn ought to be given as much importance as achievement data. At the high school level, achievement must be interpreted in the context of enrolment. Canada's performance, in light of the these concerns, appears better than some critics have suggested.

Les programmes de tests internationaux en mathématiques et en sciences requièrent une interprétation judicieuse. Les principales difficultés que posent ces tests sont, entre autres, la sous-structure de leur contenu, la signification des scores composites, l'évaluation des possibilités d'apprentissage et le choix du contenu des tests. Pour être valables, les interprétations doivent tenir compte des limites inhérentes aux comparaisons internationales, en raison par exemple des divers systèmes de valeurs, des différences entre les pays et des écarts entre les taux de scolarité. On devrait accorder autant d'importance aux possibilités d'apprentissage qu'aux données relatives au rendement scolaire. Au niveau de l'école secondaire, le rendement doit être interprété en fonction des taux de scolarité. La performance du Canada, à la lumière de ces considérations, semble meilleure que certains critiques donnent à penser.

The published results of comparative studies in mathematics and science across various countries may help both academics and laypersons to understand important educational issues. But those studies are very complex, and the task of interpreting the results requires extraordinary care (as their authors often suggest). Although the academic community, in general, is aware of the difficulties in interpreting those results, some members of the public are not so cautious. Great disparities exist between the original studies (and related specialist literature) and the media and political responses to those studies. This gap between academic and public views is a cause for concern.

There are two reasons for this gap. Some politicians are blind to issues that do not suit their purposes. "We cannot escape," Burstein (1993b) contended, "the ideological use and misuse of cross-national data for political purposes. We can

only hope to overwhelm the most base misrepresentations with the wealth of knowledge and understanding international studies provide" (p. xxxi). In addition, the media often do not take a balanced view of this type of educational data (Bracey, 1994, 1996). On the other hand, the reports themselves may be partially responsible for misinterpretation: they are often written in a language impenetrable to non-specialists.

In assessing international reports, I separate carefully the curriculum provided to students (a feature known as the "opportunity-to-learn" [OTL]), from the actual results of international tests written by students (the "achievement data"). Because curricula vary greatly across countries, and coverage of content within tests is never uniform, judgements based exclusively on achievement data, without taking into account OTL, may be unfair. Although those researchers who report the results of achievement test data are aware of this distinction, commentators who try to interpret the researchers' findings often ignore it.

STUDIES IN MATHEMATICS AND SCIENCE

The International Association for the Evaluation of Educational Achievement (IEA) has conducted two mathematics and two science studies, and is currently engaged in a third study of both. (They have also done studies in other subject areas.) The studies of mathematics and science are:

(a) First International Mathematics Study (FIMS) — a test of 13-year-olds and students in the last year of secondary school in 12 countries (Husén, 1967).
(b) First International Science Study (FISS) — a test of 10-year-olds, 14-year-olds, and students in the last year of secondary school in 19 countries (Comber & Keeves, 1973).
(c) Second International Mathematics Study (SIMS) — a test of 13-year-olds and students in the last year of secondary school in 22 countries (Burstein, 1993a; Robitaille & Garden, 1989; Travers & Westbury, 1989).
(d) Second International Science Study (SISS) — a test of 10-year-olds, 14-year-olds, and students in the last year of secondary school in 23 countries (IEA, 1988; Keeves, 1992; Postlethwaite & Wiley, 1992; Rosier & Keeves, 1991).
(e) Third International Mathematics and Science Study (TIMSS) (in progress) — a test of 9-year-olds, 13-year-olds, and those in the final year of secondary school, in approximately 50 countries (Robitaille et al., 1993).

The International Assessment of Education Progress (IAEP) has conducted two studies, the first covering both science and mathematics in one volume, and the second devoting a separate volume to each subject:

(f) First IAEP study (IAEP-1) — a test of 13-year-olds in six countries (including seven Canadian jurisdictions) (Lapointe, Mead, & Phillips, 1989).

(g) Second IAEP study (mathematics) (IAEP-2M)—a test of 9-year-olds and 13-year-olds in 21 countries (including 14 different Canadian jurisdictions) (Lapointe, Mead, & Askew, 1992).
(h) Second IAEP study (science) (IAEP-2S)—a test of 9-year-olds and 13-year-olds in 21 countries (including 14 different Canadian jurisdictions) (Lapointe, Askew, & Mead, 1992).

All of the studies cited have collected data in addition to achievement results. As well, some "countries" are, in fact, smaller jurisdictions.

My main focus is on SIMS, SISS, and the two IAEP studies; Canada was not involved in FIMS or FISS, and TIMSS staff have released preliminary results just as this article is going to press. Some problems raised concerning SISS and SIMS are being addressed by TIMSS, but my primary concern is difficulties in the interpretation of existing written reports and the influence of those reports on policy.

DIFFICULTIES IN TEST CONSTRUCTION AND DESIGN

There is wide agreement that the results of IAEP-1, for example, are questionable. "Because of the small sample size and acknowledged methodological problems," Rotberg (1990) claims, "this assessment was labelled a 'pilot'—although this label has not been reflected in the public rhetoric about the results" (p. 298). McLean (1990) refers to it as "badly flawed" (p. 10) and similarly, Goldstein (1993) states:

> Despite a caveat in the introductory section, there is little in this same report which tries to convey the tentative nature of international comparisons, and the problems of translation and interpretation which are well recognised by those responsible for designing and analysing the assessments. (p. 19)

The major difficulty with the IAEP-1 results may be traced to a difference in purpose between it and the IEA studies. Noting the time and money needed for content definition and item development in IEA studies, the IAEP-1 study investigated the feasibility of re-using NAEP (National Assessment of Educational Progress) questions and procedures (hence the label "pilot"). Existing NAEP items were selected and submitted for OTL judgements to the nine jurisdictions that agreed to participate. Although those trying to economize should not be criticized for taking such steps, their failure to be more forthcoming about limitations of the "quick and efficient" methodology led to the complaints by Goldstein (1993), McLean (1990), Rotberg (1990), and Wolfe (1989). Wolfe's analysis (see also Goldstein, 1991) demonstrates one way in which such shortcuts yielded questionable results.

Goldstein (1993) further contends that "the very notion of reporting comparisons in terms of a single scale, for example of 'mathematics' or 'science,' is

misleading" (p. 20). Providing aggregate scores, however much the public may demand them, may encourage unsound conclusions. But if authors of reports do not provide such scores, they risk having a third party produce a simplistic summary with more appeal to non-specialists. Producing digestible but not simplistic summaries of results is a difficult task that requires both effort and care. To date, only SIMS has managed to avoid the misleading single scales to which Goldstein points; the others invite simplistic one-dimensional conclusions.

DIFFICULTIES WITH TEST CONTENT

In assessing an aggregate score on these tests, it is important to understand the actual make-up of the tests and the relative weighting of subtopics within tests. In IAEP-1, for example, the relative sizes of subtests determined relative weighting of subtopics in the aggregate score. The mathematics test had six subscales, ranging from 6 to 24 items: "numbers and operations" (24 items) was treated as four times as important as "relations, functions and algebraic expressions" (6 items) and three times as important as "problem solving" (8 items). But when Wolfe (1989) adjusted the relative weighting of the subtests to something more balanced, he changed the ranking of countries. In particular, the standing of the U.K., which had low OTL for the longest subtest, improved markedly when less weight was given to this subtopic. Whether the British ought to deal more with this subtopic is another issue; the analysis simply shows that the actual items do make a big difference. (L. McLean [personal communication, 1995] notes that Educational Testing Service, who conducted the IAEP studies, disputes Wolfe's conclusions, as well as objections raised in McLean [1990].)

Consider another example on the importance of content weighting. When Burstein (1993b), using data from SIMS, took each of eight countries, and "constructed" retrospectively a test consisting only of items with at least 80% OTL rating by that country, and compared all countries on each test, he found the rank-order on the tests he constructed was substantially different from the rank-order on the original common test. In every case, the rank was higher for each country when the basis for test construction was 80% OTL within that country.

International differences in OTL raise a related concern. If a country decides not to teach topics that other countries do teach, then that decision should be discussed as seriously as the fact that students do not perform very well on topics they have not been taught. Indeed, this issue was the theme of the U.S. report on SIMS (McKnight et al., 1987). In Burstein's reconstructed test, the list of items for Japan was easily the longest, because that country teaches more of the mathematics on the IEA test than any other. Thus, the total SIMS test favoured Japan, a fact that was considered in interpreting all results in the SIMS report (Burstein, 1993a).

In summary, test content and OTL data are important. Changing content can change national ranking, and serious interpretation of the data must be at the

subtopic level. At the same time, political reality obliges investigators to report some sort of total scores. Consequently, the challenge is to find the most meaningful and least misleading scores.

DIFFICULTIES IN MEASURING OPPORTUNITY TO LEARN

To measure OTL, all that is required, in principle, is a judgement such as "fully taught," "partly taught," or "not taught." But in practice such judgements are questionable, largely because countries vary so widely in their educational structures. Whereas central officials in countries with tightly prescribed curricula assess the intended curriculum for testing purposes, in other countries teacher surveys are used to identify the implemented curriculum. Such varied information is difficult to compare. (One could even argue that students are in the best position to judge OTL.)

In addition, conditions for OTL judgement have not always been ideal. Details in the information provided to judges have not been specific enough for accurate judgements. In SISS and IAEP-2, decisions were made on quite general descriptions of items, rather than the items themselves. In SISS, for example, one biology topic was "metabolism of the organism," whereas the "detail" given to the judges was "metabolism in organisms and the structural adaptations involved." This difficulty can be solved, at some cost, by collecting judgements on actual items. SIMS collected such data, and TIMSS is doing so as well (Robitaille et al., 1993).

Finally, a peculiarly Canadian issue. OTL in SISS varied considerably across provinces, making the notion of national-level OTL suspect: "The diversity in curriculum found in Canada," Crocker (1989) contends, "leads to a serious question of whether interprovincial achievement comparisons can have any meaning" (p. 33).

DIFFICULTIES IN THE CHOICE OF CONTENT FOR INTERNATIONAL TESTS

If uniformly high agreement on OTL is required, international variation precludes producing a test of any length. In order to proceed, a compromise has been struck (Plomp, 1992): some consistency in OTL, reasonable content coverage, and insistence that the data be interpreted only in light of OTL. But the results of such a compromise were soon evident. Schmidt and Valverde (1995), for example, report that in mathematics at the Grade 4–5 level (TIMSS), no topic was reported as "substantially covered" by more than 70% of the 50 or so participating countries. (This finding flies in the face of the common perception that basic mathematics is basic mathematics.)

For science, little of the modernization of science curricula that took place between FISS and SISS (e.g., Merrill & Ridgway, 1969) reached the final SISS, with the result that the test was biased against countries in which significant

curriculum reform had taken place. For example, one goal of SISS was to allow comparisons with FISS, conducted twelve years earlier. SISS began with the 53 FISS content categories, but despite tremendous upheaval in science education in the intervening years, only four new categories could be agreed on. Crocker (1989) notes that there is rarely disagreement on what to put in a test; the issue, rather, is what to leave out. Traditional content is accepted much more readily than newer topics.

Theisen, Achola, and Boakari (1983) note that if an education system is geared to turning out "predetermined labor quotas" (p. 63) through national exams, it will have a more standardized curriculum. In that event, comparisons with more flexible systems will be misleading. One system may be geared to factual knowledge and test-taking skill, and another to student self-selection into areas of interest and ability.

Finally, attempts to examine growth by comparing different age groups on the same items do not speak well for the curricular validity of the test. SISS attempted such a comparison, with the result that the 14-year-old population wrote a core test of 30 items, all but two of which were judged suitable for and administered to either the 10-year-old or the end-of-secondary-school sample as well.

DIFFICULTIES CAUSED BY DIFFERENT SOCIAL CONDITIONS IN PARTICIPATING COUNTRIES

Even among relatively similar countries, achievement data should be interpreted in light of social and economic differences. For example, countries differ in their patterns of immigration, with the result that some have more varied language mixes than others. Those in which the language spoken at home is different from that used in school tend to score relatively poorly on achievement tests (Elley, 1992).[1] Canada has one of the highest proportions of such students (Robitaille & Garden, 1989; Rosier & Keeves, 1991).

Although social variables do not constitute an excuse for poor results, they are a major contextual feature. Jaeger (1992) recently interpreted performance on international tests (mostly FIMS and SIMS) of the United States, Germany, and Japan, in light of varying socio-economic conditions among the three countries. He notes that from 30% to 60% of achievement variance can be "predicted by the poverty rate among children in single-parent households" (p. 122), and cites similar figures for the influence of divorce rates and part-time employment on achievement. He then provides comparative data for the three countries, such as the percentage of children in single-parent families: 25% in the U.S.A. (with a 50% poverty rate), 14% in Germany (with a 36% poverty rate), and 6% in Japan (no poverty figure was given for Japan). He also analyzes the relative influence of different variables on achievement, concluding: "economic factors, coupled with family structure and stability, predict substantial portions of between-nation variation in math and science test scores . . . classroom instructional variables

predict but trivial portions" (p. 124). Even without Canadian data, Jaeger's analysis is relevant.

Nevertheless, despite these factors, poor educational achievement is seen (by some) as entirely the fault of the schools. Jaeger observes, for example, that there is more debate in the U.S.A. over being behind a half-dozen countries in achievement than there is about considerably poorer medical ratings (e.g., being ranked 28th best in percentage of low-birth-weight children). Although society recognizes poor health as a symptom of broader societal problems, and doctors are not blamed, it does not recognize poor school achievement as rooted in similar problems.

DIFFICULTIES IN THE LANGUAGE OF TEST ITEMS

A further limitation in international studies is that the tests are written in a wide variety of languages to accommodate participating countries. There are inherent difficulties crossing languages and cultures that no amount of effort or expense can eliminate. Most tests were written first in English, then translated, and finally checked by means of back-translations. Although translation procedures are improving, languages do not map onto each other without problems (Goldstein, 1993): a word in one language may differ in precise meaning and level of abstractness in another, with the result that item difficulty may vary across languages. Some differences are more cultural than linguistic — for example, some currency systems do not use decimals, so that an arithmetic item concerning, say, $1.45, cannot be translated into a corresponding item of similar difficulty in all currencies. Although translation difficulties do not account for large systematic differences in achievement, they do raise concerns about levels of error in the data.

DIFFICULTIES IN SCHOOL ENROLMENT PATTERNS

Secondary school achievement data should be interpreted in light of enrolment and retention. Beyond the age of compulsory schooling, countries differ enormously in the proportion of the age cohort in school, in the proportion who take mathematics or science, and in the types of schools they attend. According to SISS, the U.S.A., Korea, Japan, and Canada, in that order, had the highest retention rates of students to end-of-high-school age. However, Korea and Japan had substantial numbers of these students in technical or vocational schools not sampled by SISS (Postlethwaite & Wiley, 1992, p. 6). Consequently, samples vary in their degree of elitism.[2] The proportion of the age group that were eligible to be sampled varied from 80% for the U.S.A. to 18% for Hungary. Canada was at 68%, Israel and Japan over 60%, and all others below 42%. Only the U.S. sample was less elite than the Canadian.

Even more marked are the differences in proportion of students who take

advanced science. Figure 1 shows a plot of achievement versus enrolment as a proportion of the age group, in biology for 18-year-olds. Failing to consider enrolment differences results in such anomalous comparisons as between the 45% of 18-year-old Finnish students who take biology and the 5% in Hungary, England, or Singapore. Although such comparisons are clearly inappropriate, they occur in many interpretations of international data for an age when school attendance and subject selection are optional. English Canada[3] has either the highest (chemistry) or second-highest (biology, physics) enrolment of the age group in SISS, and, thus, low achievement compared to more elite systems. The SISS authors offer cautions on this point, but fail to heed their own cautions in all major analyses.

SISS produced a secondary analysis of comparable small percentages of elite students. The authors clearly state that the analysis is an examination of the effect of teaching in high-achieving homogeneous groups (Postlethwaite & Wiley, 1992, p. 70), but the data have been misinterpreted by the Economic Council of Canada (as cited in Freedman, 1993, p. 12) as achievement *adjusted for* retention rate. A comparison of one country's very best students taught in homogeneous groups with another country's best taught in heterogeneous groups is an examination of streaming. To claim that this comparison "adjusts" achievement for retention — in effect, that all other factors are equal — is to miss the point. How students are grouped is *the* major factor; it has not been equalized in this comparison.

Postlethwaite and Wiley (1992) conclude that something may be gained by homogeneous grouping of elite students. But their argument raises the question of whether it is better to achieve excellent results with a small group, or to settle for lower achievement with a larger group. If Canada's economic problems are caused by the lack of large numbers of well-trained workers for a modern economy, then surely the issue is not how well the top 5% of students are educated, but rather how successfully the top half or more of all students are educated.

Fail to consider retention and enrolment when examining secondary achievement data does a major disservice to secondary students and educators.

DIFFICULTIES CAUSED BY SAMPLING AND PARTICIPATION RATES

The differences across countries in how seriously they participate in international studies cannot be ignored. Data in international studies are collected by sampling, and agreement to participate varies across countries, generally from high in totalitarian countries to low in democracies. Schools preoccupied with their own upcoming and more important testing, such as external examinations, have tended to decline the invitation.

It is difficult to interpret the effect on comparative results of these differences across countries. Schmidt, Wolfe, and Kifer (1993) report that only 30% of the original targeted sample for SIMS in the U.S.A. agreed to participate, and the

rest of the sample was made up of "similar" schools. According to Jaeger (1992), commenting on the 71% and 74% response rates in the U.S.A. for FIMS, "response rates of the U.S. samples were below the threshold that is regarded as adequate" by the National Center for Educational Statistics (p. 119). If 75% response rate is the criterion, then one-quarter of the countries, including Canada, would be eliminated from SISS.

Differences in participation across countries may reflect national interest or pride in doing well. This interpretation is supported by comments in the IAEP-2S report, with respect to Korea: "The feeling of self-discipline and serious attention to what they are about carries over into the assessment activity . . . at this age, 12, 13, 14, students are expected to be responsible for their own serious behavior" (Lapointe, Askew & Mead, 1992, p. 24). Compare this attitude with the much more relaxed expectations Canadians have of young teens. Although these difficulties are impossible to quantify, they form part of the context for interpreting international differences.

DIFFICULTIES IN TEST AND SCORE ACCURACY

In general, levels of error reported for international tests are underestimates, and the claimed level of accuracy is an overestimate. The procedure used for estimating error rightly treats the students tested as a sample of the population of students in the country, but it does not make similar assumptions for the test itself. (Items chosen should be considered to be a sample from the population of items they are intended to represent.) Although the complexity of the negotiations required to choose items precludes the use of any standard sampling plan for the items, failure to do so directly effects error estimates.

As an example of the importance of accurate error estimates, Draper's (1995) re-analysis of published school achievement rankings in Britain is enlightening. When appropriate error levels were considered, Draper found, 15% of schools could be described as "clearly better," another 15% were "clearly poorer," and the remaining 70% were "somewhere in the middle of a large grey area" (p. 133). Nothing more precise could be justified. The situation is similar with international studies, a fact that is masked when simple rank orders are examined. Consequently, many scores close to each other should be considered "tied."

The fact that scores may be reported in different ways influences the interpretation of results. If countries are ranked by average score, then the sizes of differences are masked. Although the difference between fourth and tenth place is often trivial, this triviality is not evident in the reporting. All the studies mentioned, except SIMS, may be criticized for this failure, although it is impossible to prevent someone who is determined to produce simplistic rank orders from doing so.

If results are reported as percentages, then the brevity of many tests is masked

and small differences exaggerated. (It is also, somewhat unfortunately, very convenient for comparisons.) If two countries differ merely by one item on a 25-item test, a difference expressed as "4%" appears large. The 1988 preliminary report of SISS has received much more publicity than the 1991 and 1992 final reports, probably because of the former's relative brevity and longer time of availability. This preliminary report, however, is based on tests consisting of only 24 and 30 items.

A serious problem in reporting scores concerns the development of scaled scores using item response theory. Item-response-theory analysis can itself be controversial, because of its technical complexity (Traub & Wolfe, 1981), and because the effect of that complexity is to lose the audience (Goldstein, 1993). Those problems aside, such scaled data appear susceptible to dubious interpretations. In IAEP-1, the data were scaled to a mean of 500 and standard deviation of 100, resulting in the following (approximate) proportions being imposed on the data by the shape of the normal curve: below 300, 2%; 300s, 14%; 400s, 34%; 500s, 34% 600s, 14%; and above 600, 2%. Then, characteristics of easy, medium, and hard items were identified, and were used as labels to describe achievement at levels 300 through 700.

What's wrong with this characterization? In IAEP-1 science, these associations were made:

- 300 — know everyday facts.
- 400 — understand and apply simple scientific principles.
- 500 — use scientific procedures and analyze scientific data.
- 600 — understand and apply intermediate scientific knowledge and principles.
- 700 — integrate scientific information and experimental evidence.

Although these labels suffice as descriptors to distinguish easier from harder science content, it is not appropriate to define a scale so that, internationally, precise proportions of students should have their achievement so characterized. According to Jaeger (1992), a similar attempt at this kind of interpretation for NAEP was declared a failure by the U.S. General Accounting Office.

Another problem arises if the focus is on elite students, either by examining a given percentage of high achievers (SISS) or by setting a definition of "excellent," and comparing the numbers of students reaching this level (SIMS). Arising from nothing more than the shape of the normal curve, this type of comparison greatly exaggerates differences; the higher the standard, the greater the exaggeration. Using SISS data, I found that a 5% difference at the mean translated into a 30%–40% difference in the proportion of students exceeding a high "standard." The Economic Council of Canada (1992) reported SIMS results this way, making Canada's relative standing look much worse than that of countries with mean achievement marginally higher than Canada's.

Although achievement and retention are both important, they should be kept separate. For example, a quick reading of the SIMS data for 13-year-olds shows

the U.S.A. doing no better than one impoverished developing country. On closer reading, virtually all of the U.S. age cohort is in school, compared to only about one-third for the developing country. A suggested adjustment, to multiply the achievement by the percentage enrolment, makes the comparison more reasonable, but at the price of confounding enrolment and achievement. This calculation shows the same "effectiveness" if enrolment increases 10% and achievement decreases 10%, and it can be misinterpreted to blame schools for failing to teach those who are not in attendance. Achievement of 65% by 80% of the age group is not the same as 80% achievement by 65% of the age group.

TO WHAT EXTENT CAN THE RESULTS FOR CANADA BE ACCEPTED?

IAEP-1

This study is seriously flawed and has been heavily criticized. The authors used U.S. items, arbitrary subtest weighting, and no OTL data. The results should be ignored.

IAEP-2

This study, an improvement over IAEP-1, involved 15 countries (3 of them limited to only one language group) and 2 more geographically smaller units. There are four data sets, mathematics and science for 9- and 13-year-olds. The OTL data are weak, being derived from accounts of school administrators who reported school emphasis in very broad categories (e.g., "plants," "animals"). There are, however, some interesting classroom, home, and individual data. In three of four comparisons, Canada is tied with most countries, and ahead of one or two. The exception is mathematics for 9-year-olds, in which Canada did better, coming ahead of six countries. Canada is behind Korea in all four achievement comparisons, behind Taiwan in three of four, and behind Hungary in one.

SISS—Elementary

Although the SISS study, involving 23 countries, is much better than either IAEP study, it is hampered by OTL data based on vague descriptions of curriculum content. For age 10, the OTL data indicate great international variability in curricula; this variability is less at age 14. Assessment of the findings is hampered by the existence of several different reports, reflecting Canada's language and jurisdictional problems, and also a preliminary and final report. The original reports have engaged too much in horse-race comparisons, despite their authors' warnings.

English Canada, excluding English Quebec, is behind three countries (Korea, Finland, Japan) at age 10, tied with two, and ahead of most. At age 14, the same

Canadian group is behind Hungary, Japan, and the Netherlands, tied with five or six countries, and ahead of a similar number. English Quebec (Connelly, Crocker, & Kass, 1987) is about 5% below the national average at age 10, and 2% above at age 14. Results for French Canada are 4%–5% lower than for English Canada.

SISS—End of High School

This study is hampered by a somewhat small number of items. OTL data are uniformly quite high (that reported for chemistry in Canada is an error). I have excluded the French Canadian results, because the sample is largely Quebec Grade 11, whereas the English sample is two-thirds Grade 12 and one-third Ontario Grade 13. Vastly differing proportions of students take the subjects:

- Biology (see Figure 1): proportions of students enrolled are much higher for Canada than for all countries except Finland. Canadian achievement is in the lower third of countries. It is as high as in Australia, Sweden, and Japan, who have two-thirds the enrolment. Four countries perform about 30% better with one-quarter the enrolment; and four others perform 20% better with one-third the enrolment.
- Chemistry: no other country has even two-thirds the enrolment of Canada. Three countries perform roughly 30%–40% better with one-third the enrolment and about seven more perform 20%–30% better with one-sixth to two-thirds Canada's enrolment. Five countries perform no better, even with far lower proportional enrolment.
- Physics: Canada is second to Norway and tied with Italy in enrolment. One cluster of five countries with about half Canada's enrolment performs 25% better, and a second cluster with about two-thirds the enrolment performs 15% better. Hong Kong does about 50% better with about half Canada's enrolment.

Data exist for high-achieving constant proportions of students, taught in circumstances of greatly varying class homogeneity. But the analysis of that data exaggerates differences, and the estimate of error cannot be calculated.

- Biology: for the top 3% of the age group, English Canada is in a middle cluster of most countries, with three somewhat ahead, and three behind.
- Chemistry: for the top 5% of the age group, differences across countries are very large. Canada is very nearly in the middle, with only five countries close. England, Japan, and Hong Kong are well ahead, and Norway, Finland, and Korea well behind.
- Physics: for the top 4% of the age group, Canada's placement is similar to its placement in chemistry: Japan and Hong Kong are well ahead, whereas Italy, Finland, and Hungary are well behind. The rest (nine countries) cluster in the middle.

SIMS

The SIMS study appears to be the best of the set discussed: it reports by subscales, has better OTL data, and includes a large number of items. There are results for 13-year-olds, and end-of-high-school. Canada is represented by British Columbia and Ontario only, reported separately. Reporting is by subtest, with OTL reported on the same graph as achievement. For the younger group, a summary in standard scores (Robitaille & Garden, 1989, p. 124) allows some comparisons without subtest detail. For example, for the 13-year-olds, Ontario had its best performance in arithmetic and poorest in algebra, whereas B.C. had its best performance in arithmetic, and poorest in geometry and measurement (a tie).

Relative to 18 other jurisdictions, and arbitrarily giving equal weight to the subtests in my calculations, Ontario (Grade 13) was behind 7 countries and ahead of 7, whereas B.C. (Grade 12) was behind 4 and ahead of 13. Ontario tested Grade 12 students on most subtests used in the main SIMS project, but the Ontario results appeared in a separate study (McLean, Raphael, & Wahlstrom, 1984). On average, Ontario Grade 12 students did 15% poorer than B.C. Grade 12 students. In the main SIMS project, Japan, the Netherlands, and Hungary were the leaders. OTL differences largely parallelled the achievement differences.

For the older group, the most significant factor is the proportion of the age group in school and taking mathematics (see Figure 2). Retention and enrolment play dominant roles in achievement. For example, Hungary, which does very well at the younger level, has the lowest achievement of all jurisdictions at end-of-high-school because of its huge enrolment of the age group. No other jurisdiction comes close to Hungary's 50% — B.C. is next highest, at 30% — of the age cohort taking mathematics.

When equal weight is given to subtests, Ontario Grade 13 achievement is between that of two countries with similar enrolment, Scotland and Finland, near the middle of a group of six countries with about two-thirds Ontario's enrolment, considerably better than the U.S.A. and considerably poorer than Japan. Of the three remaining countries, with about one-third Ontario's enrolment, Ontario does more poorly than two, and slightly better than one. B.C. has the second-lowest achievement and second-highest enrolment.

SIMS also examines the achievement of the top students in constant proportions, but with differing levels of elitism in their classrooms. These results, for the top 5%, are given for three of the six subtests:

- Algebra: B.C. and Ontario are tied with two other jurisdictions, slightly behind Japan, and ahead of the others;
- Geometry: B.C. and Ontario are tied with five other jurisdictions, slightly behind Japan, and ahead of the others;
- Elementary Functions and Calculus: Ontario is tied with five countries, slightly behind Japan, and ahead of the rest; B.C. scored at the chance level, did three other countries, because calculus is not taught in that province.

In summary, results for Canada (in one case, only two provinces) for elementary age groups are fairly consistent: well in the top half or third of the countries, and usually behind two or three leaders. At end-of-high-school, interpretation is complicated by very large differences in enrolment and retention. Canada has enrolments among the highest in the world, and correspondingly lower achievement. When smaller high-achieving groups are isolated, the Canadian high school relative results appear to be fairly similar to those for elementary school, despite the fact that Canada's heterogeneous high school classes are being compared to more homogeneous groupings in other countries.

IMMEDIATE PROSPECTS

Further research is clearly needed on Canadian results in international tests. I suspect that individual and class-level socio-economic data would go a long way to explaining within-Canada variations. No variable in the control of the school system has ever been shown to be as powerful a determinant of educational achievement as student background variables.

The OTL data in international studies are as compelling as the achievement results. Although vastly different OTLs make achievement comparisons questionable, the data also prompt questions about the quality of the school curriculum. When one country gives high OTL to twice as many items as another country, it must raise the question of whether that second country has in place a defensible curriculum. On the other hand, if more mathematics or science are included in a curriculum, other subjects may suffer. To get more science or mathematics into the curriculum will require time, money, and the willingness to take something else out.

Canadian participation in international testing will undoubtedly persist. Given current public concern about education, it would be politically foolish to withdraw from such tests. Over the last few decades substantial progress has been made in both the design of international tests and the methods analyzing them (Burstein, 1993a; Goldstein, 1987; Schmidt & Valverde, 1995; Wolfe, 1989). If all countries work with the same broad set of variables, similar countries can learn from each other's within-country analyses: examining the same variables, learning from each other's mistakes, and avoiding costly dead-ends.

ACKNOWLEDGEMENT

The original report on which this article is based was written for the 1993–1994 Ontario Royal Commission on Learning. I thank Xiaofang Shen for assistance in the preparation of the original report, and Les McLean, Howard Russell, Alan Ryan, and Ross Traub for detailed comments on earlier versions of the article.

NOTES

[1] Elley's (1992) study was of reading, but the point stands for mathematics and science.

[2] Canadians may wish to debate the advisability of such schools in Canada, but that is a separate question from the comparability of achievement results.

[3] These data come from the preliminary report of SISS, published before the study of French Canada was completed.

REFERENCES

Bracey, G. W. (1994, October). The fourth Bracey report on the condition of public education. *Phi Delta Kappan, 76,* 115–127.

Bracey, G. W. (1996). International comparisons and the condition of American education. *Educational Researcher, 25*(1), 5–11.

Burstein, L. (Ed.). (1993a). *The IEA study of mathematics III: Student growth and classroom processes.* New York: Pergamon Press.

Burstein, L. (1993b). Prologue. In L. Burstein (Ed.), *The IEA study of mathematics III: Student growth and classroom processes* (pp. xxvii–lii). New York: Pergamon Press.

Comber, L. C., & Keeves, J. P. (1973). *Science education in nineteen countries.* New York: John Wiley.

Connelly, F. M., Crocker, R. K., & Kass, H. (1987). *Second international science study: Phase II.* Unpublished Report to the Social Sciences and Humanities Research Council of Canada.

Crocker, R. K. (1989). *Science achievement in Canadian schools: National and international comparisons.* Unpublished Report on the Second International Science Study prepared for the Economic Council of Canada.

Draper, D. (1995). Inference and hierarchical modelling in the social sciences. *Journal of Educational and Behavioral Statistics, 20,* 115–147.

Economic Council of Canada. (1992). *A lot to learn: Education and training in Canada.* Ottawa: Author.

Elley, W. B. (1992). *How in the world do students read?* Hamburg, Germany: Grindeldruck GMBH.

Freedman, J. (1993). *Failing grades: Redirecting Canada's educational debate.* Red Deer, AB: Full Court Press.

Goldstein, H. (1987). *Multilevel models in educational and social research.* New York: Oxford University Press.

Goldstein, H. (1991). More thoughts on testing. *Education Canada, 31*(1), 44–47.

Goldstein, H. (1993). *Interpreting international comparisons of student achievement.* Report prepared for UNESCO, Paris.

Husén, T. (1967). *International study of achievement in mathematics: A comparison of twelve countries* (Vols. 1–2). New York: Wiley.

International Association for the Evaluation of Educational Achievement [IEA]. (1988). *Science achievement in seventeen countries: A preliminary report.* New York: Pergamon Press.

Jaeger, R. M. (1992, October). World class standards, choice, and privatization: Weak measurement serving presumptive policy. *Phi Delta Kappan, 74,* 118–128.

Keeves, J. P. (Ed). (1992). *The IEA study of science III: Changes in science education and achievement, 1970 to 1984.* New York: Pergamon Press.

Lapointe, A. E., Askew, J. M., & Mead, N. A. (1992). *Learning science.* Princeton, NJ: Educational Testing Service.

Lapointe, A. E., Mead, N. A., & Askew, J. M. (1992). *Learning mathematics.* Princeton, NJ: Educational Testing Service.

Lapointe, A. E., Mead, N. A., & Phillips, G. W. (1989). *A world of differences: An international assessment of mathematics and science.* Princeton, NJ: Educational Testing Service.

McKnight, C. C., Crosswhite, F. J., Dossey, J. A., Kifer, E., Swafford, J. O., Travers, K. J., & Cooney, T. J. (1987). *The underachieving curriculum: Assessing U.S. school mathematics from an international perspective.* Champaign, IL: Stipes Publishing.

McLean, L. (1990). Let's call a halt to pointless testing. *Education Canada, 30*(3), 10–13.

McLean, L., Raphael, D., & Wahlstrom, M. (1984). *Intentions and attainments in the teaching and learning of mathematics: Report on the Second International mathematics study in Ontario, Canada.* Toronto: Ontario Institute for Studies in Education, Educational Evaluation Centre.

Merrill, R. J., & Ridgway, D. W. (1969). *The CHEMStudy story: A successful curriculum improvement project.* San Francisco: Freeman.

Plomp, T. (1992). Conceptualizing a comparative educational research framework. *Prospects, 22,* 278–288.

Postlethwaite, T. N., & Wiley, D. E. (1992). *The IEA study of science II: Science achievement in twenty-three countries.* New York: Pergamon Press.

Robitaille, D. F., & Garden, R. A. (Eds.). (1989). *The IEA study of mathematics II: Contexts and outcomes of school mathematics.* New York: Pergamon Press.

Robitaille, D. F., Schmidt, W. H., Raizen, S., McKnight, C., Britton, E., & Nicol, C. (1993). *Curriculum frameworks for mathematics and science* (Third International Mathematics and Science Study, Monograph 1). Vancouver: Pacific Educational Press.

Rosier, M. J., & Keeves, J. P. (Eds). (1991). *The IEA study of science I: Science education and curricula in twenty-three countries.* New York: Pergamon Press.

Rotberg, I. C. (1990, December). I never promised you first place. *Phi Delta Kappan, 72,* 296–303.

Schmidt, W. H., & Valverde, G. A. (1995, April). *Cross-national and regional variations in mathematics and science curricula.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Schmidt, W. H., Wolfe, R. G., & Kifer, E. (1993). The identification and description of student growth in mathematics achievement. In L. Burstein (Ed.), *The IEA study of mathematics III: Student growth and classroom processes* (pp. 59–99). New York: Pergamon Press.

Theisen, G. L., Achola, P. P. W., & Boakari, F. M. (1983). The underachievement of cross-national studies of achievement. *Comparative Educational Review, 27,* 46–68.

Traub, R. E., & Wolfe, R. G. (1981). Latent trait theories and the assessment of educational achievement. *Review of Research in Education, 9,* 377–435.

Travers, K. J., & Westbury, I. (1989). *The IEA study of mathematics I: Analysis of mathematics curricula.* New York: Pergamon Press.

Wolfe, R. G. (1989, February). *An indifference to differences: Problems with the IAEP-88 study.* Paper presented at a National Science Foundation-sponsored conference on the Second International Mathematics Study data at the University of Illinois, Champaign, IL.

Philip Nagy is a Professor in the Measurement and Evaluation program of the Department of Curriculum, Teaching, and Learning of the Ontario Institute for Studies in Education of the University of Toronto, 252 Bloor Street West, Toronto, Ontario, M5S 1V6.